RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback

Year: 2023 | Citations: 474 | Authors: Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard

Abstract

Reinforcement learning from human feedback (RLHF) has proven effective in aligning large language models (LLMs) with human preferences, but gathering high-quality preference labels is expensive. RL from AI Feedback (RLAIF), introduced in Bai et al., offers a promising alternative that trains the reward model (RM) on preferences generated by an off-the-shelf LLM. Across the tasks of summarization, helpful dialogue generation, and harmless dialogue generation, we show that RLAIF achieves comparable performance to RLHF. Furthermore, we take a step towards"self-improvement"by demonstrating that RLAIF can outperform a supervised fine-tuned baseline even when the AI labeler is the same size as the policy, or even the exact same checkpoint as the initial policy. Finally, we introduce direct-RLAIF (d-RLAIF) - a technique that circumvents RM training by obtaining rewards directly from an off-the-shelf LLM during RL, which achieves superior performance to canonical RLAIF. Our results suggest that RLAIF can achieve performance on-par with using human feedback, offering a potential solution to the scalability limitations of RLHF.